Airbnb Room Price Prediction

Xiaofei Wu

I. Introduction

1. Project overall

This report analyze an Airbnb dataset to predict the price (in $US) for a night stay. through Linear mixed effect model with log transformation. Before all model were build, the exploratory data analysis (EDA) is used to analyzing data sets to summarize their main characteristics, mostly with visual methods. The model will be evaluated from the root mean squared error (RMSE), where y hat is the natural log of the predicted price for (in $US) for a night stay and y is the natural log of the actual price value. Linear mixed effect model with log transformation leads to the least of 0.365945. Then, to share the results from analysis, the report will contain the interpretation. In the end, limitation and future work are talked to for future analysis and improvement.

2. Background

The data is collected by Tom Slee from the public Airbnb web site without logging in. Tom Slee's Report: How Airbnb hid the facts in New York City inspired me to looking into the Airbnb room price in New York City and see if I can find the pattern of price setting for Airbnb hosts.

3. Evaluation

The goodness of fit will be evaluated on the root mean squared error. RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

4. **Data source**

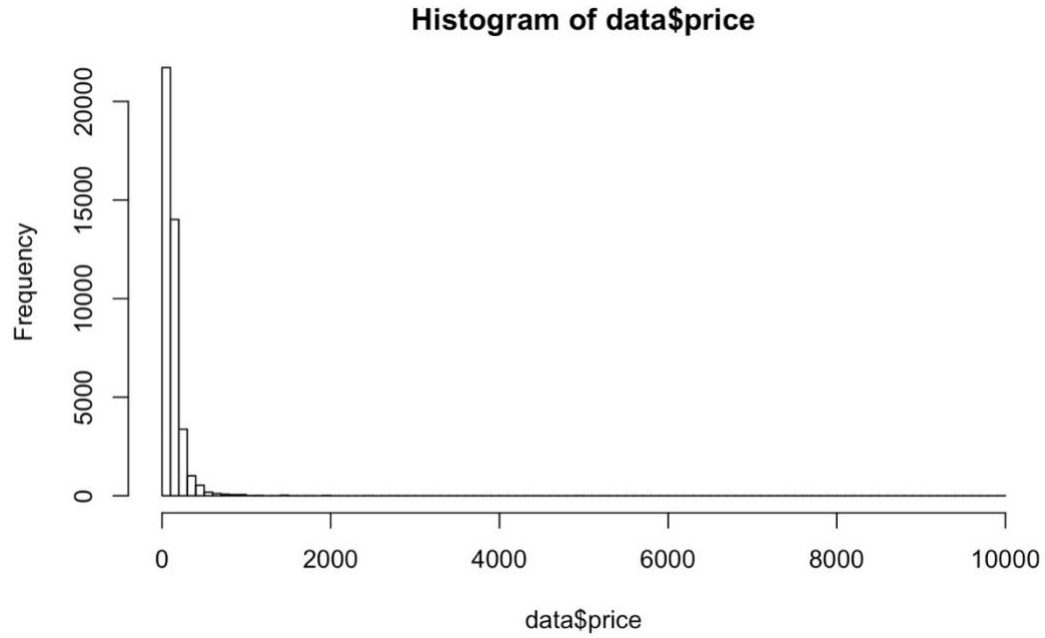
The data set is all from Tom Slee's website (To learn more about the data source, please visit <http://tomslee.net/airbnb-data-collection-get-the-data>).

My analysis contains 'room_id', 'host_id' to identify an Airbnb host, 'room_type' as One of "Entire home/apt", "Private room", or "Shared room". 'Borough', 'neighborhood'(a subset of the city that is smaller than a borough), 'reviews'(number of reviews from the room), 'overall satisfaction' (a star rate higher indicating more satisfaction), 'accommodates'(the number of guests a listing can accommodate), 'bedrooms', the price (in $US) for a night stay, 'latitude' and 'longitude'.

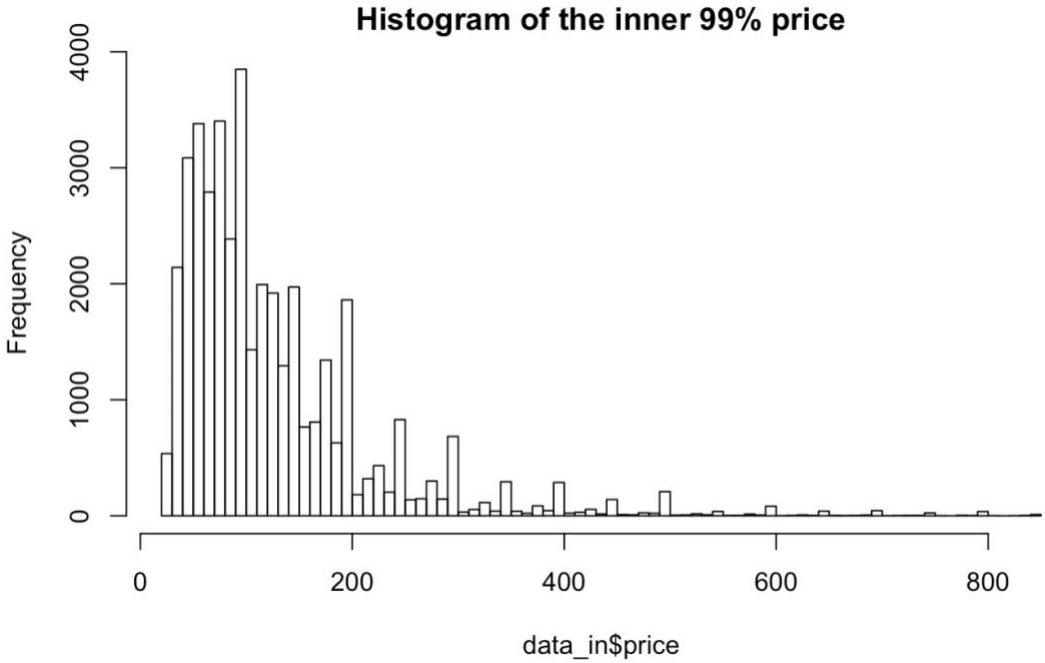
I delete columns for country and city since it is all from New York, as well as columns with null value such as 'minstay' and 'bathroom'.

II. Exploratory Data Analysis(EDA)

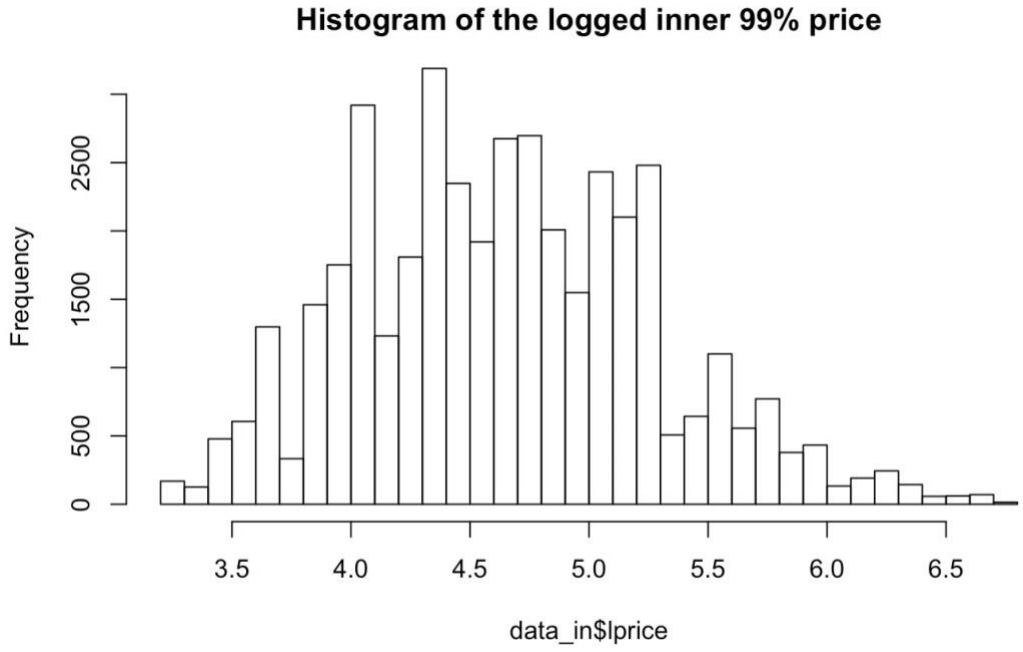
1. Firstly group the data by prices, which is the outcome, and create plots.



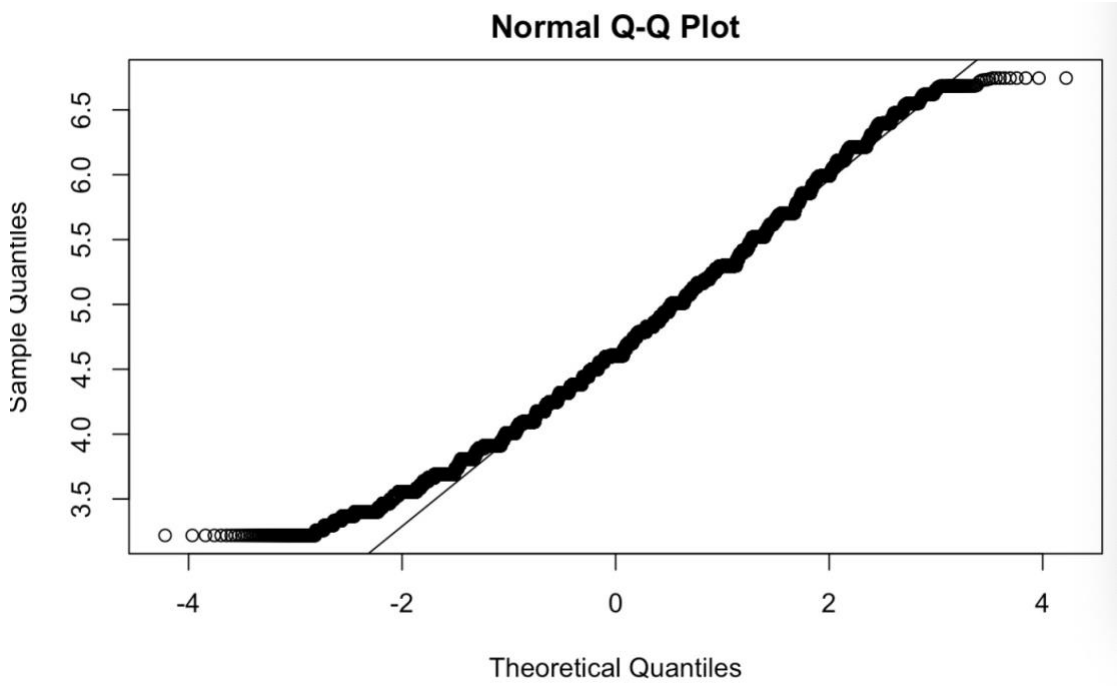
The prices are not normally distributed, with max price 9999 and min price 0. I want to avoid extreme price values and use the inner 99% price for prediction.



I took log transformation on the price and get a distribution with max 6.745236 and min 3.218876. The minimum price for price in the middle 99% is 25, so I avoided 0.



It seems like a normal distribution, so I created QQ plot to check:



It indicates that the inner 99% target after log transformation approximately follows normal distribution.

2. **Analyze the correlations between explanatory and response variables**

First I want to see the correlations between all numerical variables.

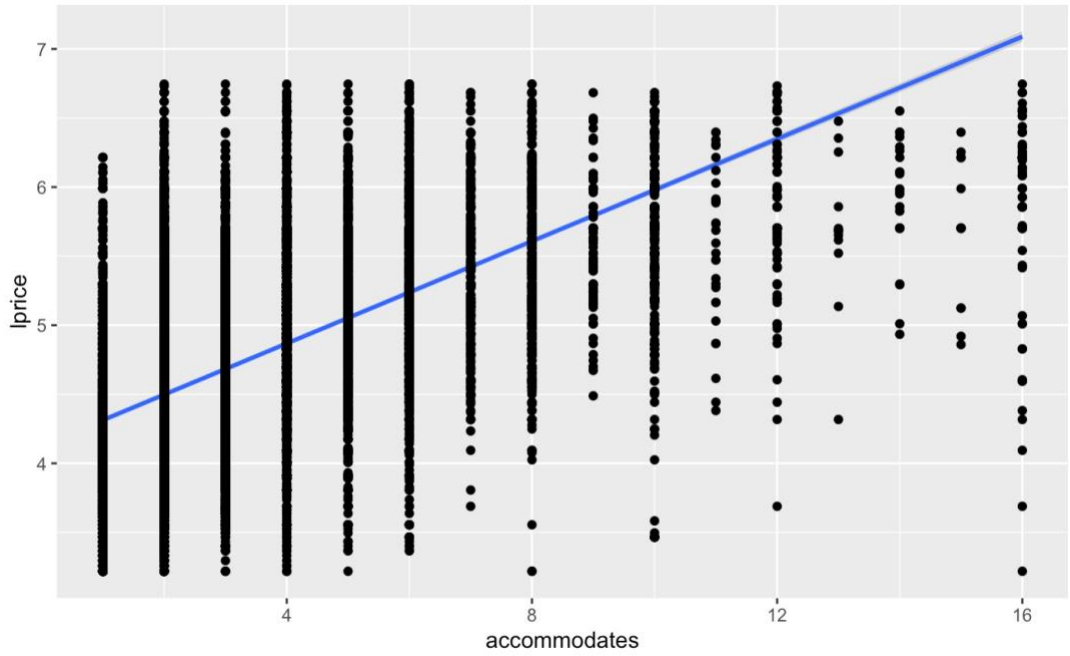
	price	reviews	overall_satisfaction	accommodates	bedrooms	latitude	longitude
price	1.00000000	-0.027169524	-0.011692089	0.54043211	0.38597307	0.043254775	-0.267561090
reviews	-0.02716952	1.000000000	0.441406701	0.10845397	0.02405399	-0.001915253	0.000833890
overall_satisfaction	-0.01169209	0.441406701	1.000000000	0.11550660	0.01487418	-0.010960273	-0.003369888
accommodates	0.54043211	0.108453970	0.115506596	1.00000000	0.62179426	-0.053825185	-0.023086648
bedrooms	0.38597307	0.024053994	0.014874183	0.62179426	1.00000000	-0.073628434	0.020445663
latitude	0.04325477	-0.001915253	-0.010960273	-0.05382519	-0.07362843	1.000000000	0.097482071
longitude	-0.26756109	0.000833890	-0.003369888	-0.02308665	0.02044566	0.097482071	1.000000000

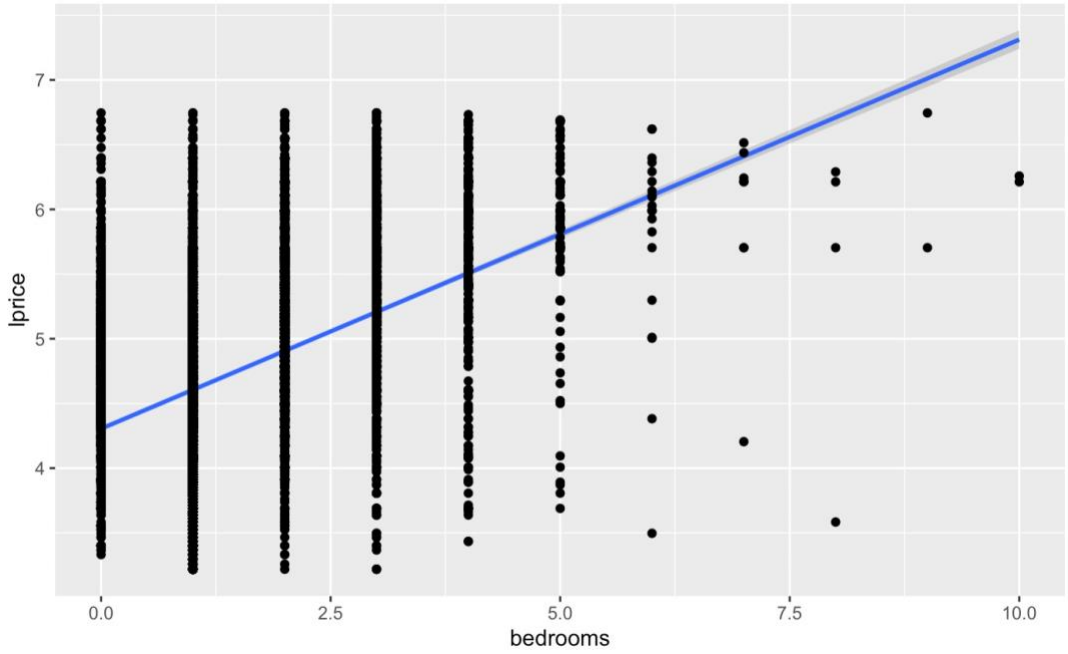
From the table, accommodates and bedrooms have apparently positive correlations with price, it makes sense that bigger room has higher price. Longitude has negative correlation with price, I guess it is caused by different overall room price in different boroughs. Overall satisfaction has positive correlation with review number.

3. **Analyze from the Room characteristic: “room_type”, "accommodates", "bedrooms","property_type".**

   a) accommodates & bedrooms

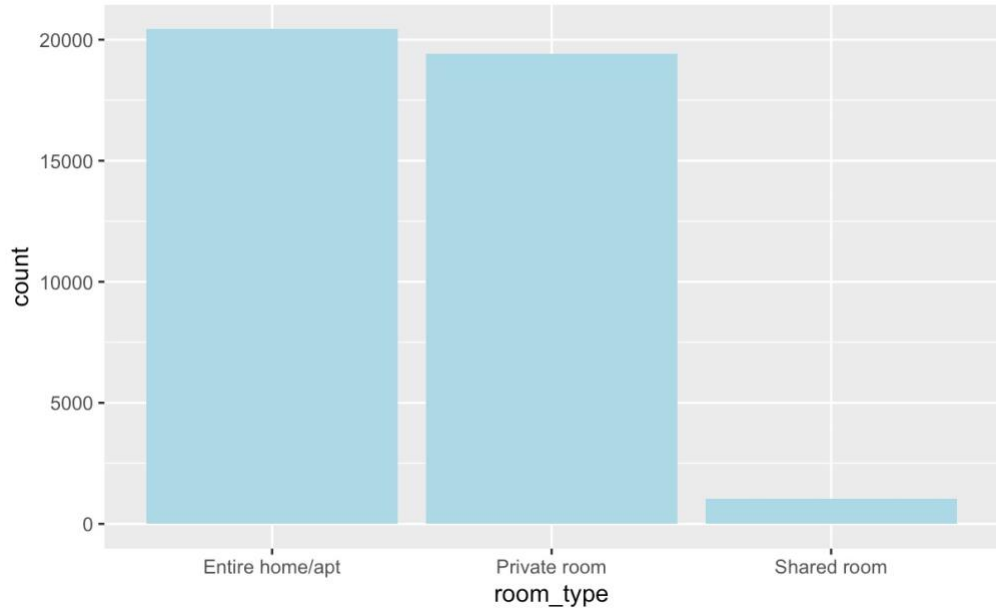
   There is one room with 50 bedrooms, I delete it and draw the following plots.

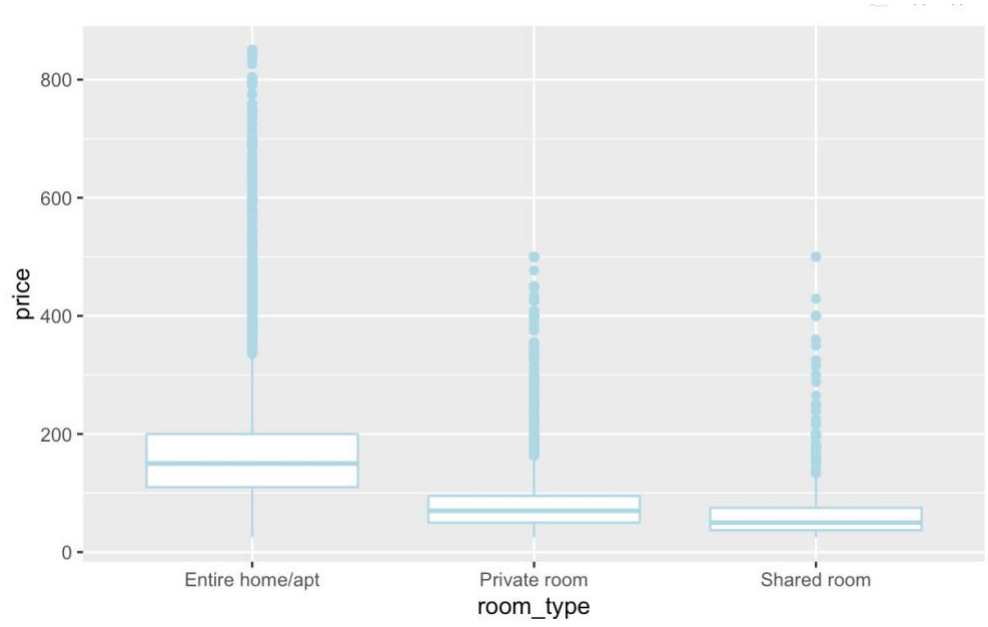




CONCLUSION: There is linear correlation between price and accommodates or bedrooms, consider accommodates and bedrooms as fixed effects. Adding (accommodates + bedrooms) in the linear mixed models. However, the price has a wide range for any fixed accommodates and bedrooms value, some random effect should be considered towards to these two fixed predictors.

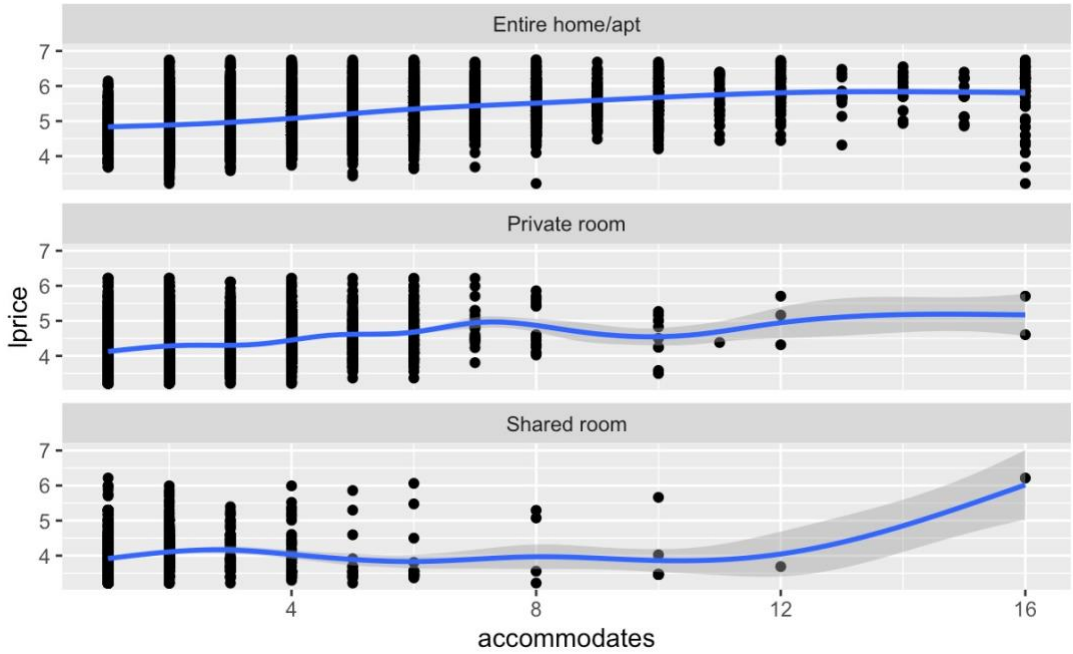
b) room_type

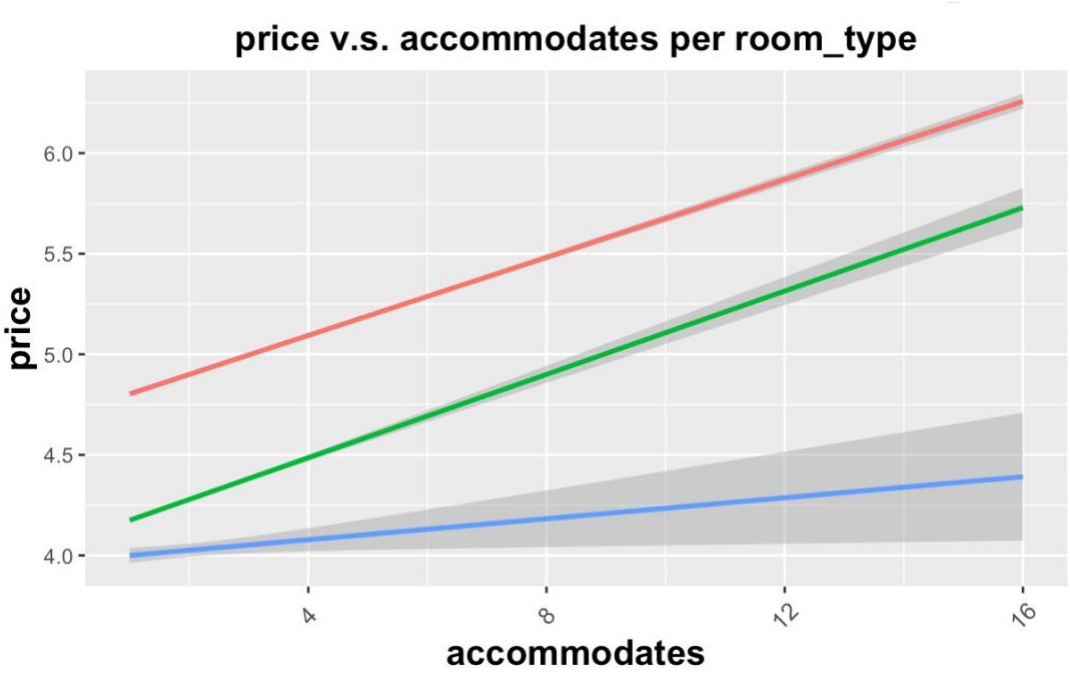




CONCLUSION: different room_type has different price distribution, consider room_type in the model.

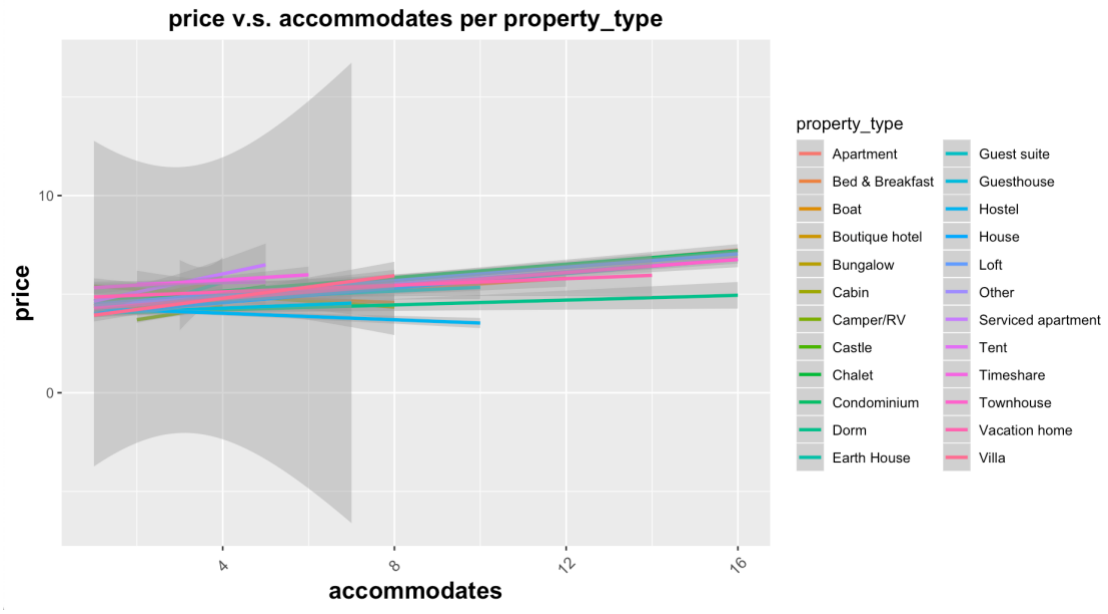
c) Room types & accommodates:





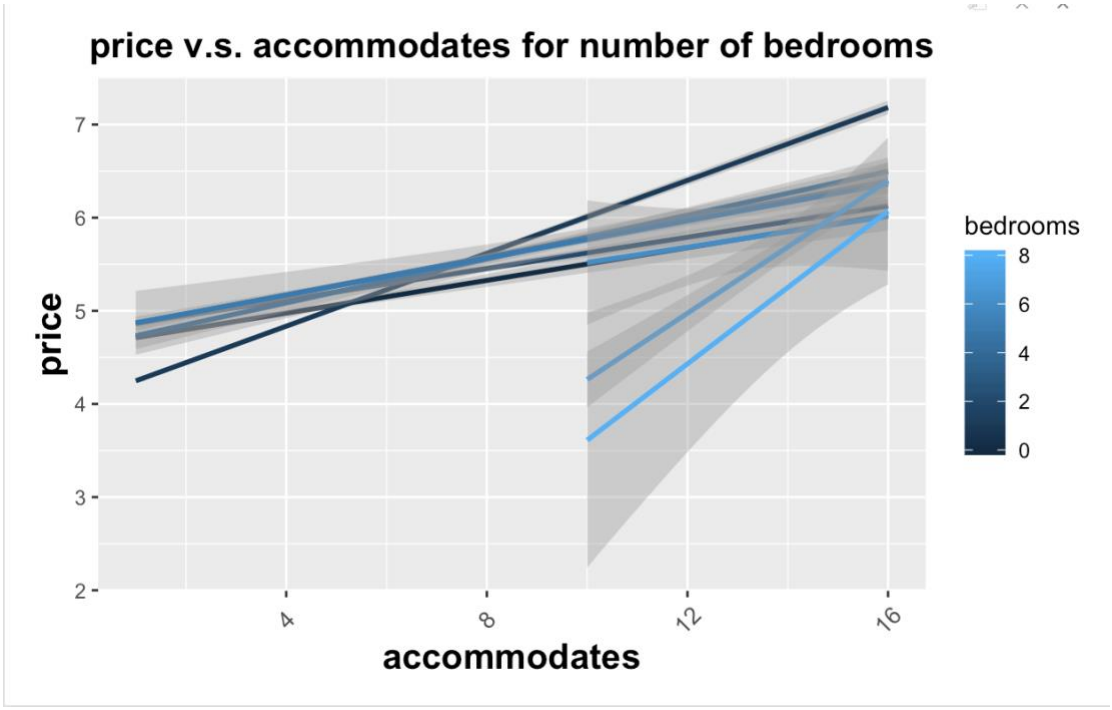
CONCLUSION: The effect (slope) of accommodates on price differ from room_type, Adding (1+accommodates|room_type) in the model.

d) Property_type & Accommodates



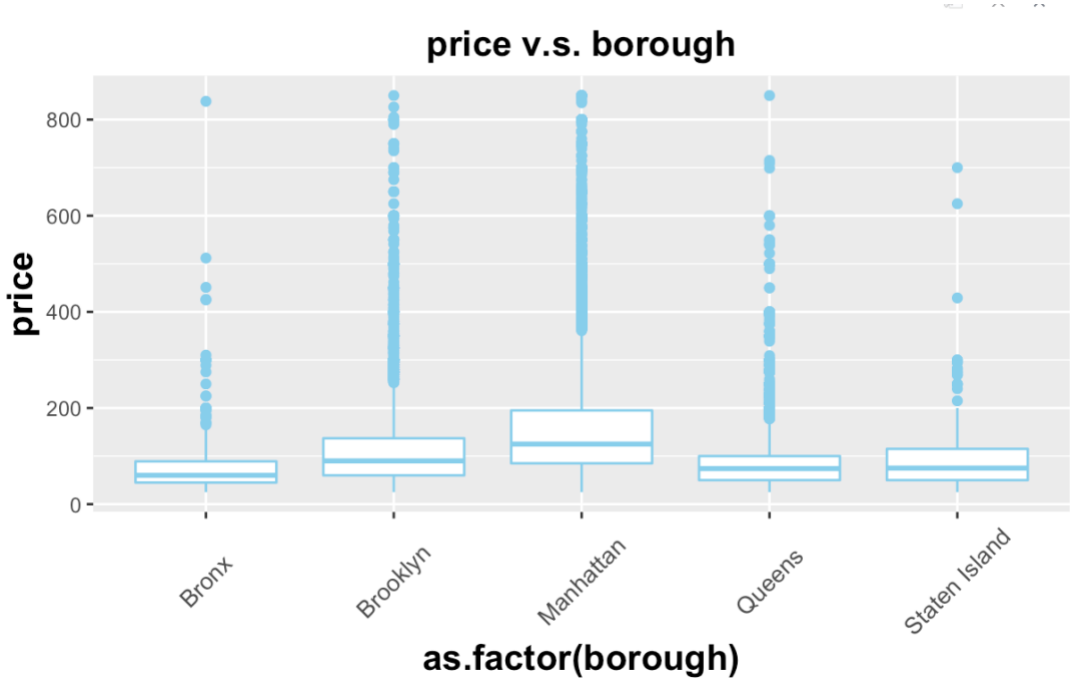
CONCLUSION: No significant effect (slope) of accommodates on price differ from Property_type.

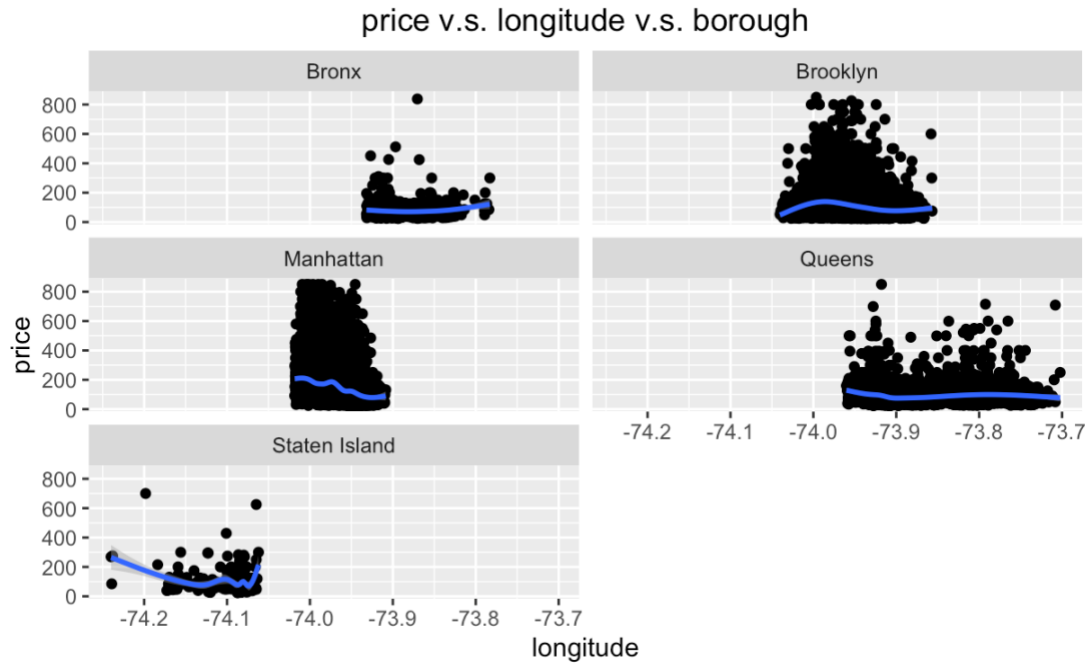
e) Bedroom & Accommodates



CONCLUSION: The effect (slope) of accommodates on price differ from bedrooms, but we should either add bedroom as a fixed effect or as a random effect (1+accommodates|bedrooms) in the model.

4. Analyze geometrical: borough, neighborhood, latitude, longitude





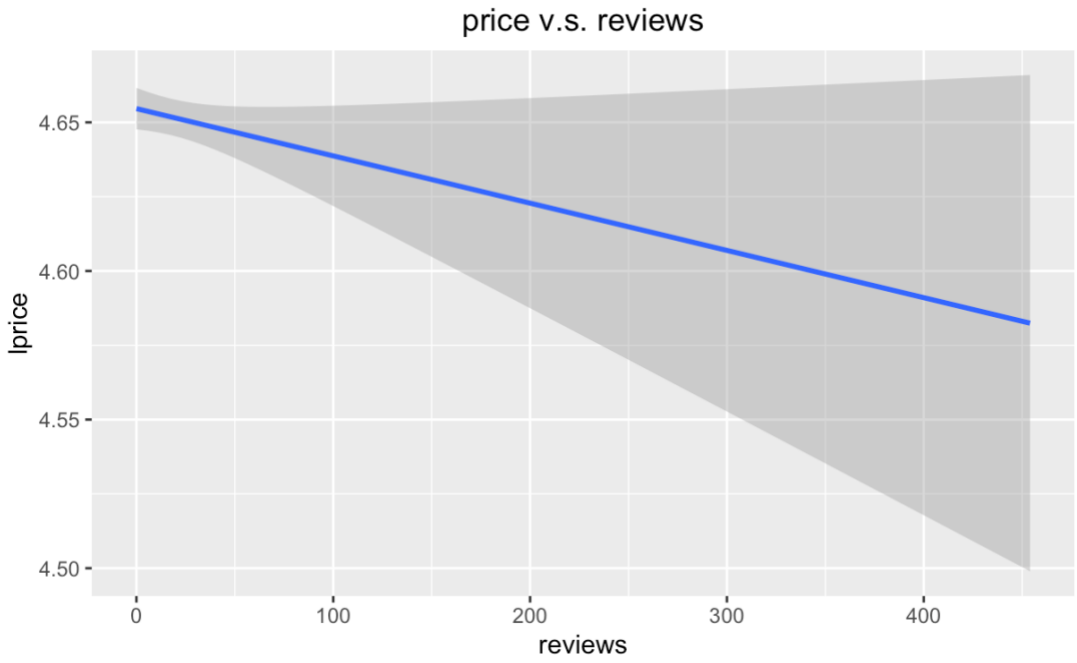
CONCLUSION: My assumption "borough causes the correlation between borough and price" seems to be right. The longitude for boroughs differ a lot, the slopes of longitude towards price did not show any signal or differ by boroughs. Different boroughs has apparent different price.

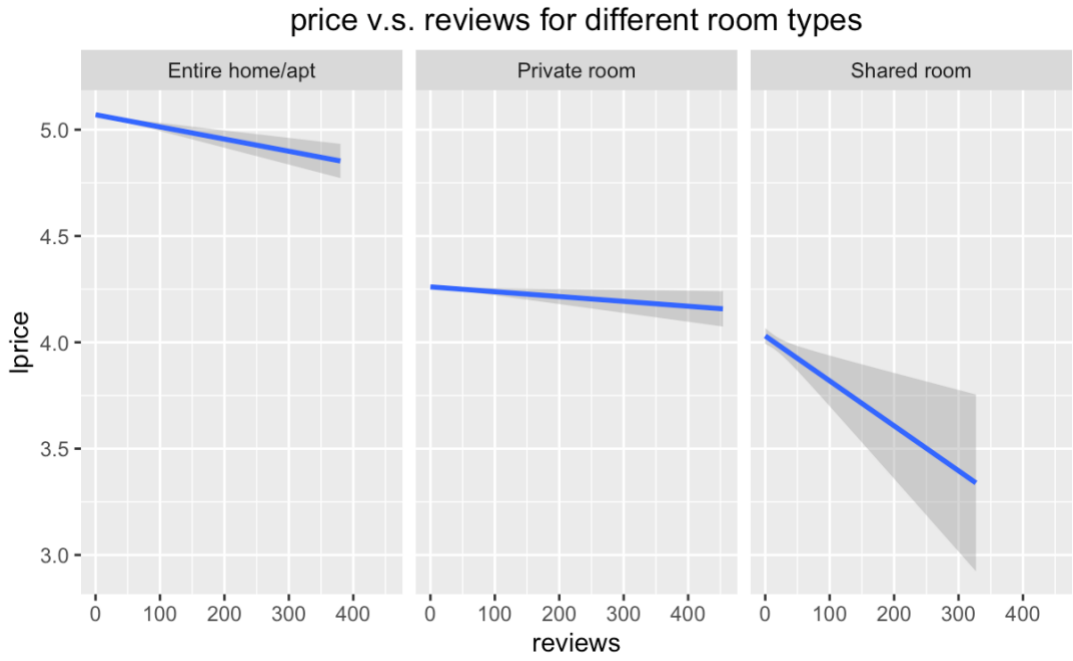
Tabulate the neighborhood with the average price：

Bronx	Brooklyn	Manhattan	Queens	Staten Island
75.86657	110.75733	156.54179	90.50153	96.85455

5. Analyze review: reviews, overall satisfaction

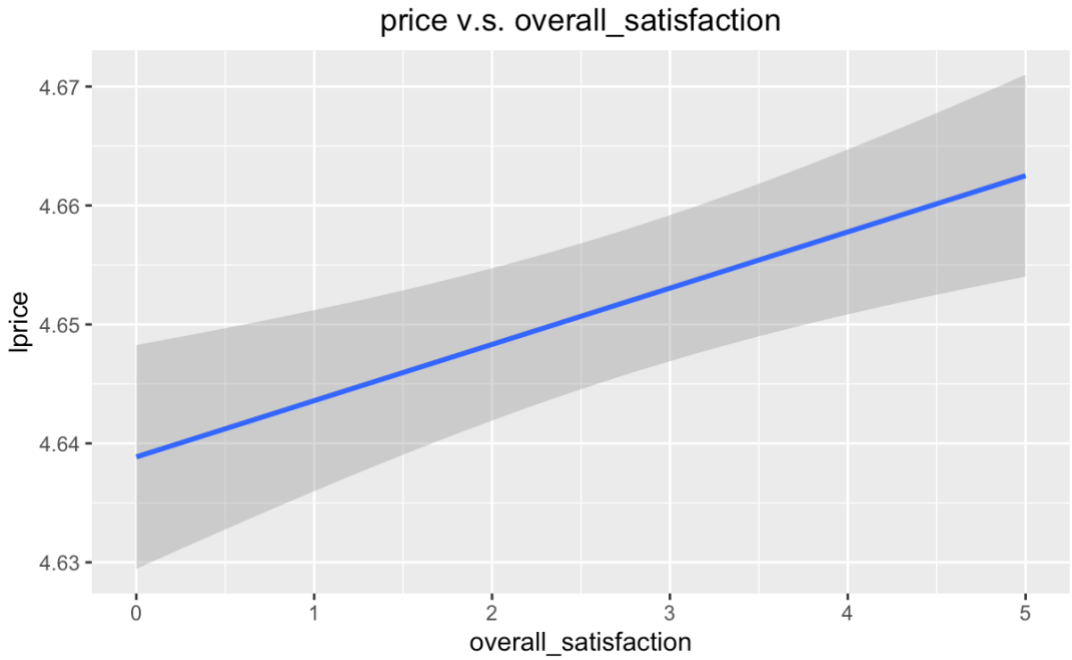
a) reviews

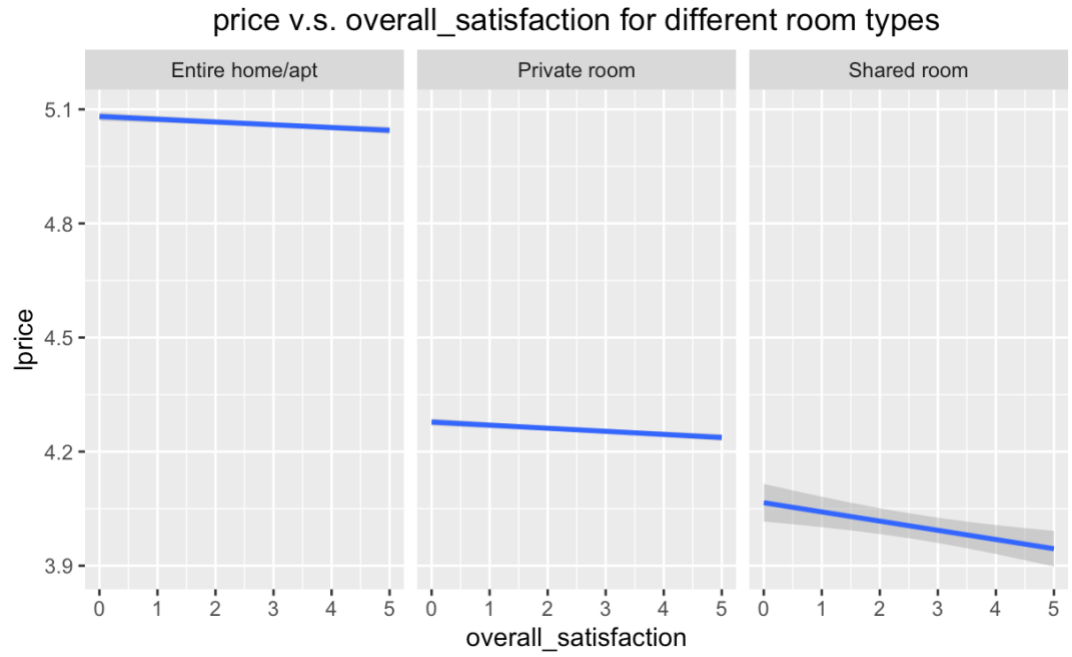




CONCLUSION: more reviews imply lower price in general, for various accommodates and various room types.

b) overall satisfaction





CONCLUSION: Higher overall satisfaction imply higher price in general, however, for any specific room types, the slope of overall satisfaction are not very significant, which means when we groups outcome by room type, it is questionable to say the overall satisfaction could imply the price.

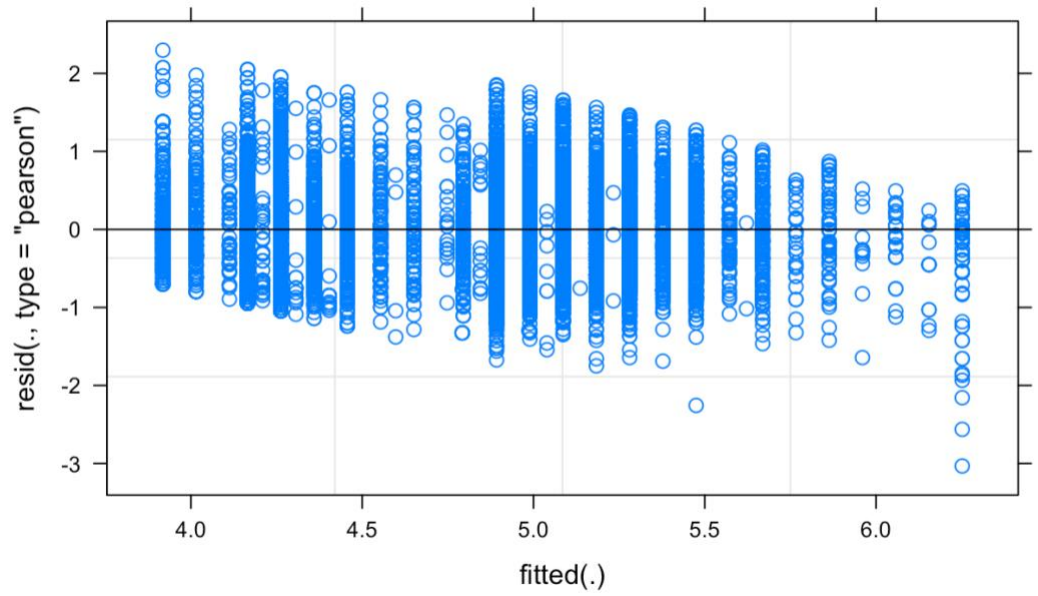
III. Model

1. Fit linear mixed model group by room types. [RMSE = 0.4524676]

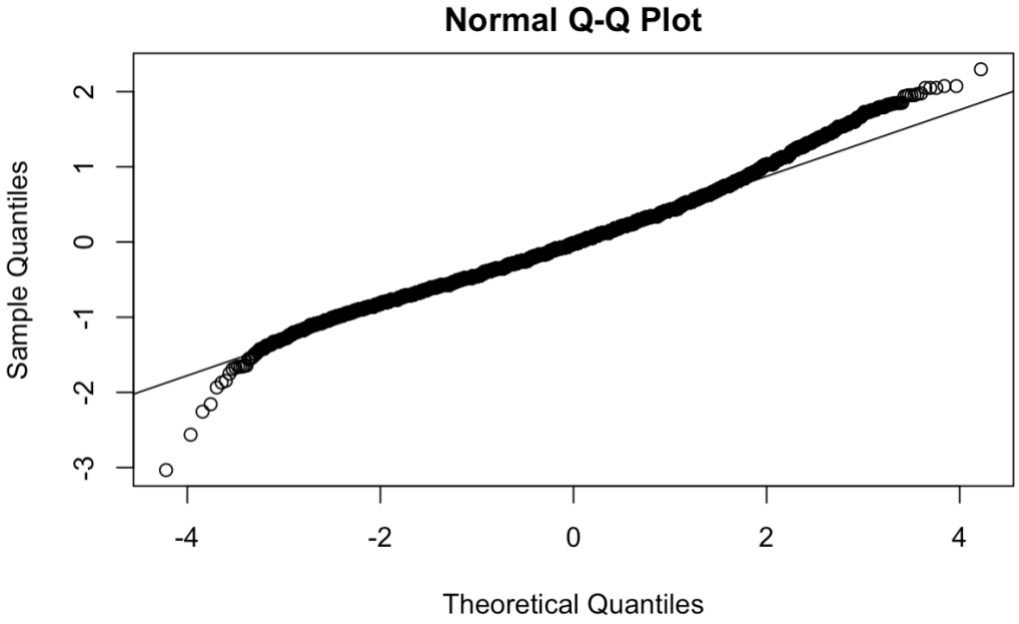
```
Linear mixed model fit by REML ['lmerMod']
Formula: lprice ~ accommodates + (1 | room_type)
   Data: data_in

REML criterion at convergence: 51221.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7025 -0.6792 -0.0491  0.6379  5.0751

Random effects:
 Groups    Name        Variance Std.Dev.
 room_type (Intercept) 0.2046   0.4523
 Residual              0.2047   0.4525
Number of obs: 40886, groups:  room_type, 3

Fixed effects:
             Estimate Std. Error t value
(Intercept)  4.195815   0.261206   16.06
accommodates 0.097100   0.001421   68.31

Correlation of Fixed Effects:
            (Intr)
accommodats -0.014
```

The goodness of fit will be evaluated on the root mean squared error. RMSE is 0.4524676.

Check residuals using plot(MD1) to see if the **linear mixed model** follows assumptions (1. The errors have constant variance. 2. The errors are independent. 3. The errors are Normally distributed.**)**



Residuals decrease on the right side.

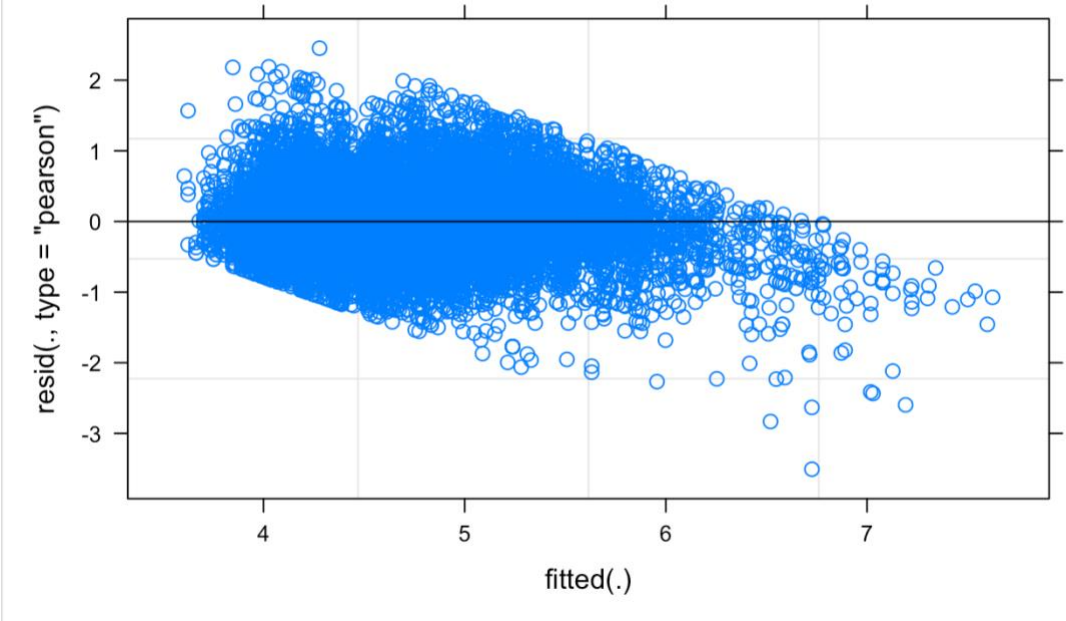


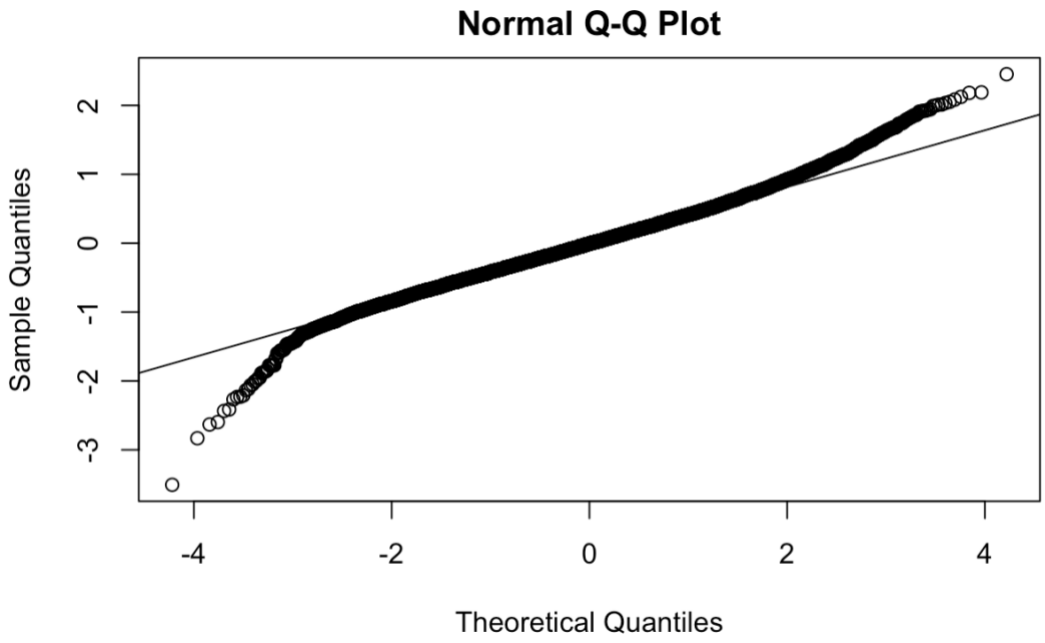
QQ plot show the residual follows the normal distribution except the two ends.

2. **Fit linear mixed model group by neighborhood. (including the effect of borough) [RMSE = 0.4356327]**

```
Linear mixed model fit by REML ['lmerMod']
Formula: lprice ~ accommodates + (1 | borough:neighborhood)
   Data: data_in

REML criterion at convergence: 48949.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.0315 -0.6534 -0.0157  0.6193  5.6162

Random effects:
 Groups                Name        Variance Std.Dev.
 borough:neighborhood (Intercept) 0.09964  0.3157
 Residual                         0.19069  0.4367
Number of obs: 40886, groups:  borough:neighborhood, 233

Fixed effects:
             Estimate Std. Error t value
(Intercept)  3.950258   0.022899   172.5
accommodates 0.179235   0.001206   148.6

Correlation of Fixed Effects:
            (Intr)
accommodats -0.154
```

Check Residuals:





QQ plot show the residual follows the normal distribution except the two ends.

3. **Fit linear mixed model group by both neighborhood and room types. (including the effect of borough) [RMSE = 0.3715812]**

```
lmer(formula = lprice ~ accommodates + bedrooms + reviews + (1 +
    accommodates | room_type) + (accommodates | borough:neighborhood),
    data = data_in)
            coef.est coef.se
(Intercept)  4.01     0.24
accommodates 0.08     0.02
bedrooms     0.12     0.00
reviews      0.00     0.00

Error terms:
 Groups               Name         Std.Dev. Corr
 borough:neighborhood (Intercept)  0.23
                      accommodates 0.03     0.21
 room_type            (Intercept)  0.41
                      accommodates 0.03     -0.43
 Residual                          0.37
---
number of obs: 40886, groups: borough:neighborhood, 233; room_type, 3
AIC = 36273.1, DIC = 36180.6
deviance = 36215.8
```

4. **Fit linear mixed model group by both neighborhood and room types, adding more fixed variables. (including the effect of borough) [RMSE = 0.3675805]**

```
lmer(formula = lprice ~ accommodates + bedrooms + reviews + overall_satisfaction +
    property_type + latitude + longitude + (1 + accommodates |
    room_type) + (accommodates | borough:neighborhood), data = data_in)
                              coef.est coef.se
(Intercept)                   -195.50    17.07
accommodates                     0.08     0.01
bedrooms                         0.11     0.00
reviews                          0.00     0.00
```

…

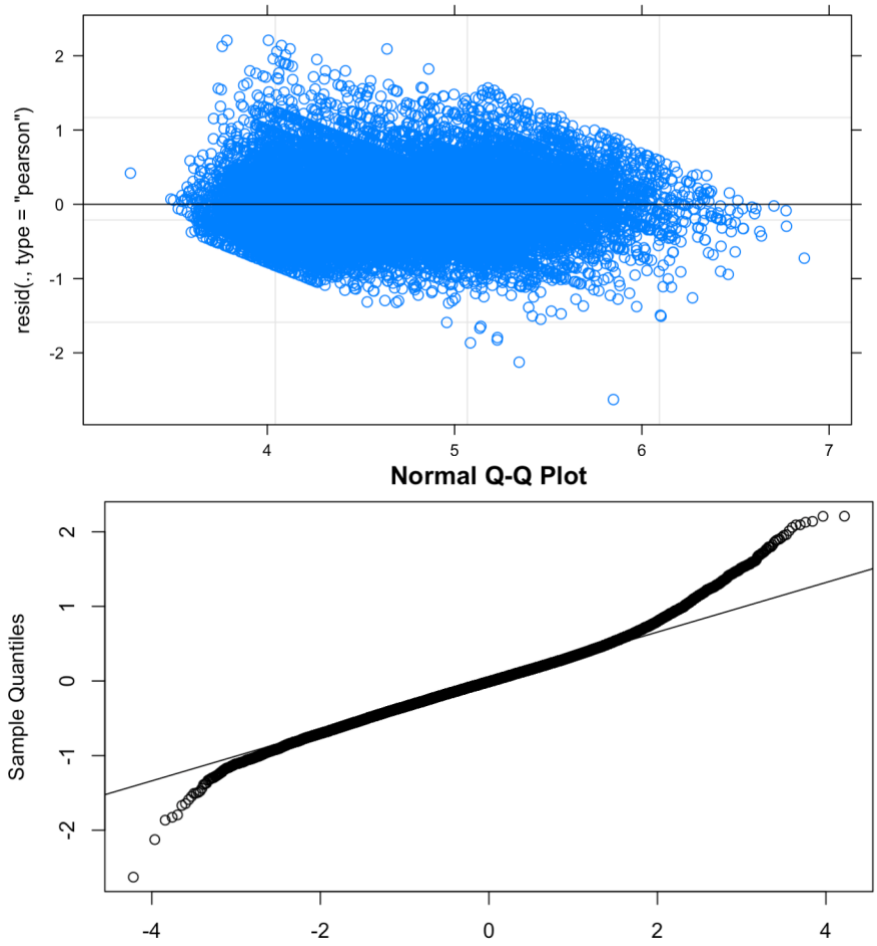
```
Error terms:
 Groups                Name         Std.Dev. Corr
 borough:neighborhood (Intercept)  0.28
                      accommodates 0.03     0.13
 room_type            (Intercept)  0.42
                      accommodates 0.02     -0.54
 Residual                          0.37
---
number of obs: 40886, groups: borough:neighborhood, 233; room_type, 3
AIC = 35626.4, DIC = 35280.3
deviance = 35412.3
```

5. **Fit linear mixed model group by both neighborhood, room types and bedrooms, adding more fixed variables. (including the effect of borough) [RMSE = 0.365945]**

```
lmer(formula = lprice ~ accommodates + reviews + overall_satisfaction +
    property_type + latitude + longitude + (1 + accommodates |
    room_type) + (accommodates | borough:neighborhood) + (accommodates |
    bedrooms), data = data_in)
                              coef.est coef.se
(Intercept)                   -198.25    17.14
accommodates                     0.09     0.02
reviews                          0.00     0.00
overall_satisfaction            -0.01     0.00
property_typeBed & Breakfast     0.22     0.03
```

```
Error terms:
 Groups                Name         Std.Dev. Corr
 borough:neighborhood (Intercept)  0.28
                      accommodates 0.03     0.19
 bedrooms             (Intercept)  0.18
                      accommodates 0.01     0.45
 room_type            (Intercept)  0.43
                      accommodates 0.03     -0.64
 Residual                          0.37
---
number of obs: 40886, groups: borough:neighborhood, 233; bedrooms, 12; room_type, 3
AIC = 35279.4, DIC = 34948.8
deviance = 35071.1
```

Check residuals:



IV. Interpretation

linear mixed model group by both neighborhood, room types and bedrooms.

Predictors	+/-	estimate	t value
accommodates	+	8.602e-02	4.941
reviews	-	7.918e-04	-12.298
overall_satisfaction	-	8.328e-03	-9.475
longitude	+	2.534e+00	-14.699

Random effects	Intercept/Slope to accommodates	Variance
borough:neighborhood	**Intercept**	**0.0810395**
bedrooms	**Intercept**	**0.0324233**
room_type	**Intercept**	**0.1887975**
Residual		**0.1349951**

Total random variance: 0.435

The best fit model use fixed predictors: accommodates + reviews + overall_satisfaction + property_type + latitude + longitude, grouped by room types, boroughs, neighborhoods, and bedrooms.

From the tables, we can see that the most significant predictor in the final model is accommodates, reviews, overall satisfaction and longitude. Larger accommodates and longitude determined higher price, on the other side, higher reviews or overall satisfaction imply lower price. This result roughly follows my assumptions in the EDA.

From the perspective of random effects, both room types and borough:neighborhood explain a lot of the variance.

V. Discussions:

1. Assessment of the result

Linear mixed effect model with log transformation leads to the least of 0.365945. The outcome gives me some explanations on the pricing for Airbnb in New York city.

2. Limitations

There are more data about Airbnb pricing online, due to time concerns, I only used this dataset. Also, as I know, there are some professional managers that help hosts to manage the house, for different type of managers, there could be pattern exists when they pricing the room. I did not get these information from this dataset. However, due to time concerns, I will keep this thoughts for future.

3. Future directions

a) Time: The date and time could be considered in the model if I got the corresponding data. For example, some Airbnb hosts raise price on weekend, lower the price if the room is not booked late that day. Date and time could be a huge impact on the final transaction price.

b) **Multiple Dataset：** there is a great website <http://insideairbnb.com/> contains more dataset about Airbnb. Also, it contains some sophisticated research on Airbnb data through which I could get deeper understanding of the Airbnb business. Then consider new perspective for this project.

VI. Acknowledgement

I would like express many thanks to Professor Masanao Yajima for helping me gaining all these modeling knowledge.

VII. Reference

1. Data source: http://tomslee.net/category/about
2. Tom Slee's Report: How Airbnb hid the facts in New York City: http://insideairbnb.com/how-airbnb-hid-the-facts-in-nyc/